Please type a plus sign (+) inside this box → [+]

PTO/SB/05 (4/98)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 C.F.R. § 1.53(b))

Attorney Docket No.	501.38779X00
First Inventor or Application Identifier	Toru HISAMITSU
Title	See 1 in Addendum
Express Mail Label No.	

JC892 U.S. PTO 09/642771 08/22/00

JC535 U.S. PTO 08/22/00

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

ADDRESS TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231

1. [X] * Fee Transmittal Form *(e.g., PTO/SB/17)* *(Submit an original and a duplicate for fee processing)*
2. [X] Specification [*Total Pages* 28] *(preferred arrangement set forth below)*
   - Descriptive title of the Invention
   - Cross References to Related Applications
   - Statement Regarding Fed sponsored R & D
   - Reference to Microfiche Appendix
   - Background of the Invention
   - Brief Summary of the Invention
   - Brief Description of the Drawings *(if filed)*
   - Detailed Description
   - Claim(s)
   - Abstract of the Disclosure
3. [X] Drawing(s) *(35 U.S.C. 113)* [*Total Sheets* 8]
4. Oath or Declaration [*Total Pages* 4]
   - a. [X] Newly executed (original or copy)
   - b. [ ] Copy from a prior application (37 C.F.R. § 1.63(d)) *(for continuation/divisional with Box 16 completed)*
     - i. [ ] DELETION OF INVENTOR(S) Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

** NOTE FOR ITEMS 1 & 13 IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).*

5. [ ] Microfiche Computer Program *(Appendix)*
6. Nucleotide and/or Amino Acid Sequence Submission *(if applicable, all necessary)*
   - a. [ ] Computer Readable Copy
   - b. [ ] Paper Copy (identical to computer copy)
   - c. [ ] Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

7. [X] Assignment Papers (cover sheet & document(s))
8. [ ] 37 C.F.R.§3.73(b) Statement *(when there is an assignee)* [X] Power of Attorney
9. [ ] English Translation Document *(if applicable)*
10. [X] Information Disclosure Statement (IDS)/PTO-1449 [X] Copies of IDS Citations
11. [ ] Preliminary Amendment
12. [X] Return Receipt Postcard (MPEP 503) *(Should be specifically itemized)*
13. [ ] * Small Entity Statement(s) *(PTO/SB/09-12)* [ ] Statement filed in prior application Status still proper and desired
14. [X] Certified Copy of Priority Document(s) *(if foreign priority is claimed)*
15. [X] Other: See 2 in Addendum

16. **If a CONTINUING APPLICATION,** *check appropriate box, and supply the requisite information below and in a preliminary amendment:*

[ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No: _____/_____

Prior application information: Examiner _____ *Group / Art Unit:* _____

For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

17. CORRESPONDENCE ADDRESS

[X] *Customer Number or Bar Code Label* 020457 *(Insert Customer No. or Attach bar code label here)* *or* [ ] *Correspondence address below*

Name					
Address					
City		*State*		*Zip Code*	
Country		*Telephone*		*Fax*	

Name (Print/Type)	Carl I. Brundidge	*Registration No. (Attorney/Agent)*	29,621
Signature		*Date*	8-22-00

Burden Hour Statement This form is estimated to take 0 2 hours to complete. Time will vary depending upon the needs of the individual case Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

Attachment to PTO/SB/05 (4/98) Utility Patent Application Transmittal

1. WORD IMPORTANCE CALCULATION METHOD, DOCUMENT RETRIEVING INTERFACE, WORD DICTIONARY MAKING METHOD
2. - FIGS.1-8
   - CREDIT CARD PAYMENT FORM

PTO/SB/17 (12/99)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

FEE TRANSMITTAL for FY 2000

Patent fees are subject to annual revision.
Small Entity payments must be supported by a small entity statement, otherwise large entity fees must be paid. See Forms PTO/SB/09-12.
See 37 C.F.R. §§ 1.27 and 1.28.

TOTAL AMOUNT OF PAYMENT ($) 886.00

Complete if Known	
Application Number	
Filing Date	August 22, 2000
First Named Inventor	Toru HISAMITSU
Examiner Name	
Group / Art Unit	
Attorney Docket No.	501.38779X00

METHOD OF PAYMENT (check one)

1. ☐ The Commissioner is hereby authorized to charge indicated fees and credit any overpayments to:

Deposit Account Number: 01-2135

Deposit Account Name: Antonelli,Terry,Stout&Kraus,LLP

☒ Charge Any Additional Fee Required Under 37 CFR §§ 1.16 and 1.17

2. ☒ **Payment Enclosed:**
☐ Check ☐ Money Order ☒ Other

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
101	690	201	345	Utility filing fee	690.00
106	310	206	155	Design filing fee	
107	480	207	240	Plant filing fee	
108	690	208	345	Reissue filing fee	
114	150	214	75	Provisional filing fee	

SUBTOTAL (1) ($) 690.00

2. EXTRA CLAIM FEES

	Extra Claims		Fee from below	Fee Paid
Total Claims 9	-20** = 0	X	18	= 0
Independent Claims 5	- 3** = 2	X	78	= 156
Multiple Dependent				= 0

**or number previously paid, if greater; For Reissues, see below

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
103	18	203	9	Claims in excess of 20
102	78	202	39	Independent claims in excess of 3
104	260	204	130	Multiple dependent claim, if not paid
109	78	209	39	** Reissue independent claims over original patent
110	18	210	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($) 156.00

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
105	130	205	65	Surcharge - late filing fee or oath	0.00
127	50	227	25	Surcharge - late provisional filing fee or cover sheet.	0.00
139	130	139	130	Non-English specification	0.00
147	2,520	147	2,520	For filing a request for reexamination	0.00
112	920*	112	920*	Requesting publication of SIR prior to Examiner action	0.00
113	1,840*	113	1,840*	Requesting publication of SIR after Examiner action	0.00
115	110	215	55	Extension for reply within first month	0.00
116	380	216	190	Extension for reply within second month	0.00
117	870	217	435	Extension for reply within third month	0.00
118	1,360	218	680	Extension for reply within fourth month	0.00
128	1,850	228	925	Extension for reply within fifth month	0.00
119	300	219	150	Notice of Appeal	0.00
120	300	220	150	Filing a brief in support of an appeal	0.00
121	260	221	130	Request for oral hearing	0.00
138	1,510	138	1,510	Petition to institute a public use proceeding	0.00
140	110	240	55	Petition to revive - unavoidable	0.00
141	1,210	241	605	Petition to revive - unintentional	0.00
142	1,210	242	605	Utility issue fee (or reissue)	0.00
143	430	243	215	Design issue fee	0.00
144	580	244	290	Plant issue fee	0.00
122	130	122	130	Petitions to the Commissioner	0.00
123	50	123	50	Petitions related to provisional applications	0.00
126	240	126	240	Submission of Information Disclosure Stmt	0.00
581	40	581	40	Recording each patent assignment per property (times number of properties)	40.00
146	690	246	345	Filing a submission after final rejection (37 CFR § 1.129(a))	0.00
149	690	249	345	For each additional invention to be examined (37 CFR § 1.129(b))	0.00
Other fee (specify)					0.00
Other fee (specify)					0.00

*Reduced by Basic Filing Fee Paid

SUBTOTAL (3) ($) 40.00

SUBMITTED BY

Complete (if applicable)

Name (Print/Type)	Carl I. Brundidge	Registration No. (Attorney/Agent)	29,621	Telephone	703-312-6600
Signature				Date	8-22-00

WARNING:

Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Washington, DC 20231.

TITLE OF THE INVENTION

WORD IMPORTANCE CALCULATION METHOD, DOCUMENT RETRIEVING INTERFACE, WORD DICTIONARY MAKING METHOD

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a technique for measuring the importance of words or word sequences in a group of documents, and is intended for use in supporting document retrieval and automatic construction of a word dictionary among other purposes.

Description of the Related Art

Fig. 1 illustrates a document retrieval system having windows for displaying "topic words" in the retrieved documents, wherein the window on the right side selectively displays words in the documents displayed in that on the left side. An example of such a system is disclosed, for example, in the Japanese Published Unexamined Patent Application No. Hei 10-74210, "Document Retrieval Supporting Method and Document Retrieving Service Using It" (Reference 1).

Kyo Kageura (et al.), "Methods of automatic term recognition: A review," *Terminology*,1996) (Reference 2) describes a method for calculating the importance of words. Methods to calculate the importance of words has long been

studied with a view to automatic term extraction or facilitating literature searching by weighting words characterizing the desired document.

Words may be weighted either to extract important words from a specific document or to extract important words from all documents. The best known in connection with the former is tf-idf, where idf is the logarithm of the quotient of the division of the total number N of documents by the number N(w) of documents in which a certain word w occurs while tf is the frequency of occurrence f(w, D) of the word in a document d; tf-idf, as the product of these factors, is represented by:

$f(w, d) \times \log2(N/N(w))$

There are variations including the following square root of f(w, d):

$f(w, d)^{**}0.5 \times \log2(N/N(w))$. Whereas there further are many other variations, tf-idf is set, as its basic nature, to become "greater as the word occurs more frequently and concentrates in a smaller number of documents."

Though not stated in Reference 2, a natural method to expand this measure, instead of pertaining to the importance of a word in a specific document, into a measure of the importance of the word in the set of all documents is to replace f(w, d) with f(w), the frequency of w in all documents.

One of the methods to extract important words from all documents is to measure the accidentalness of differences in the frequency of occurrence of each word from one given document category to another, and to qualify as important words what have a higher degree of non-accidentalness. The accidentalness of differences can be measured by several measures including the chi-square test, and this method requires the categorization of the document set in advance.

In a separate context from these studies, there are a series of attempts to identify a collection of words (or word sequences) which qualify as important words (or word sequences) from the standpoint of natural language processing. In these studies, methods have been proposed by which words (or word sequences) to be judged as important are to be restricted by the use of grammatical knowledge together with the intensity of the co-occurrence of adjoining words assessed by various measures. As such measures, there are used (pointwise) mutual information, the log-likelihood ratio and so forth.

BRIEF SUMMARY OF THE INVENTION

Techniques so far used involve the following problems: (1) tf-idf (or its like) is not accurate enough - the contribution of the frequency of a word empirically tends to be too large, making it difficult to exclude such

too common stop-words as "do"; (2) while a method to compare differences in the distribution of a specific word among categories requires the classification of documents in advance, this requirement generally is not satisfied; (3) a method to utilize the intensity of co-occurrence between adjoining words cannot evaluate the importance of a single word. It is also not easy to extend the methods so that it can treat a word sequence containing *n* words (n>2); and (4) the setting of a threshold value for selecting important words has been difficult and apt to be *ad hoc*. An object of the present invention is to provide a method free from such problems.

In the following description, a "term" means a word or a word sequence. To paraphrase the "importance of a term" from the viewpoint of term extraction or information retrieval, that a given term is important means that the term indicates or represent a topic (or topics) of some significance, in other words, the term is informative or domain-specific. In the following, such a term is said to be "representative" and in this context the "importance" of a term is also called the representativeness of a term. Since such a term is likely to be useful in taking an overview of the contents of a document set, it is important in information retrieval or a support system thereto.

In measuring the degree of representativeness, a conventional method would take only the distribution of the pertinent term itself. However, a method like tf-idf is not accurate enough though simple, or a method using a statistic such as the chi square involves difficulty in obtaining statistically significant values for most of terms because the frequency of a term is too low to properly apply such statistical test, except in rare cases, and this results in low precision.

The present invention takes note not of the distribution of a specific term but of the distribution of words occurring in association with the term noted. This is based on a working hypothesis that "the representativeness of a term is related to the unevenness of the distribution of words occurring together with the term" and that a given term is "representative" means that "the distribution of words occurring with the term are characteristic."

Therefore, the present invention uses, in calculating the representativeness of a word W, the difference between the word distribution in D(W), the set of documents which consists of every document containing W, and the word distribution in the whole documents from which said D(W) derives. In particular, the characteristic consists in that the difference is determined by comparing two distances, d and d'. Here, d is the distance between said D(W) and the

whole documents, and d', the distance between a randomly selected subset of documents containing substantially the same number of words as said D(W) and the whole documents, where the concept of "distance between two documents" includes the distance between two word distributions: that in one document set and that in another.

Other and further objects, features and advantages of the invention will appear more fully from the following description.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

A preferred form of the present invention is illustrated in the accompanying drawings in which:

Fig. 1 shows an example of information retrieval support system having a window to display topic words;

Fig. 2 shows an example of distance between two word distributions;

Fig. 3 shows a hardware configuration for realizing a proposed word importance calculation method;

Fig. 4 shows the configuration of a representativeness calculation program;

Fig. 5 shows an example of configuration for use in applying representativeness to displaying of retrieved documents in support of document retrieval;

Fig. 6 shows an example of configuration for use in applying representativeness to automatic word extraction;

Fig. 7 is a graph of results of an experiment showing how the proposed word importance raises the ranks of words considered suitable for summarizing the results of retrieval in comparison with other measures;

Fig. 8 is a graph of results of an experiment showing how the proposed word importance lowers the ranks of words considered unsuitable or unnecessary for summarizing the results of retrieval in comparison with other measures.

DETAILED DESCRIPTION OF THE INVENTION

The present invention will be described in detail below.

First will be explained the signs used for implementing the invention; 301 denotes a storage; 3011, text data; 3012, a morphological analysis program; 3013, a word-document association program; 3014, a word-document association database (DB); 3015, a representativeness calculation program; 3016, a representativeness DB; 3017, a shared data area; 3018, a working area; 302, an input device; 303, a communication device; 304, a main memory; 305,a CPU; 306, a terminal device; 4011, a module for calculating background word distribution; 4012, module for calculating baseline function; 4013, a document extraction

module; 4014, a module for calculating co-occurring word distribution; 4015, a module for calculating distance between word-distributions; 4016, a module for normalizing distance between word distributions; 4017, a random sampling module; 544, a topic words displaying routine; 5441, a topic words extraction routine; 5442, a co-occurrence analysis routine; 5443, a graph mapping routine; 5444, a graph displaying routine; 601, storage devices; 6011, text data; 6012, a morphological analysis program; 6013, a word-document association program; 6014, a word-document association database; 6015, a database for extracted words; 6016, a working area; 6017, a representativeness calculation program; 6018, a representativeness DB; 6019, a shared data area; 601A, a program for extracting word sequences; 601B, a program for grammatical filtering; 601C, a filtering program; 602, an input device; 603, a communication device; 604, a main memory; 605, a CPU; and 606, a terminal device consisting of a display, a keyboard and so forth.

The following description will concern a method for assessing the representativeness of any term and its application to an information retrieval system. First, measures of assessing the representativeness of a term is introduced by mathematically rephrasing the idea stated in BRIEF SUMMARY OF THE INVENTION above. Thus, with respect

to any term W (word or word sequence), note is taken of the word distribution in D(W), the set of documents that consists of every document containing the term W and the word distribution in the whole documents. More specifically, Rep(W), the representativeness of W is defined on the basis of Dist{PD(W), P0}, the distance of two distributions PD(W) and P0, where D0 is the set of the whole documents; PD(W), word distribution in D(W); P0, word distribution in D0.

Whereas many methods of measuring the distance between word distributions are conceivable, the principal ones of which include (1) the log-likelihood ratio, (2) Kullback-Leibler divergence, (3) transition probability and (4)vector-space model (cosign method), it has been confirmed that steady results can be obtained by using, for instance, the log-likelihood ratio. The distance between PD(W) and P0, using the log-likelihood ratio, is defined below where {w1, ..., wn} represent all words, and ki and Ki, the frequencies of the occurrence of a word wi in D(W) and D0, respectively.

Numerical expression 1:

$$\sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}$$

Fig. 2 displays words corresponding to coordinates $(\#D(W), Dist\{PD(W), P0\})$s where W varies over said words, and also it plots coordinates $(\#D, Dist\{P_D, P_0\})$s where D

varies over randomly selected document sets, where the displayed words and the document sets are taken from articles in the 1996 issues of a financial newspaper *Nihon Keizai Shimbun*.

As is seen in Fig. 2, comparison of *Dist{PD(W1), P0}* and *Dist{PD(W2), P0}* is consistent with what human intuition tells when #D(W1) and #D(W2) are close to each other. For instance, "USA" has a higher value of Dist{PD(W), P0} than "*suru*" (do) and so does "Aum", which is the name of an infamous cult, than "combine". However, a pair of terms whose #D(W) values widely differ, (this means that there is a large difference between the frequency of two terms) cannot be appropriately compared in terms of representativeness, because usually Dist{PD(W), P0} increases as #D(W) increases. Actually, "Aum" and "*suru*" are about equal in Dist{PD(W), P0}, which is against human linguistic intuition. Then, in order to offset the intrinsic behavior of Dist{ $\cdot$, P0}, {(#D, Dist{PD, P0})}s plotted in Fig. 2 using "×" marks are to be investigated. These points are likely to be well approximated by a single smooth curve beginning at (0, 0) and ending at (#D0, 0). This curve will be hereinafter referred to as the baseline curve.

Whereas it is evident that by definition Dist{PD, P0} is 0 when D = ϕ and D = D0, it has been confirmed that the behavior of the baseline curve in the neighborhood of (0,

0) is stable and similar to each other when the size of the whole documents varies over a broad range(say, about 2,000 documents to a full-year total of newspapers amounting to about 3000,000 documents).

Then, an approximating function B(・) is figured out in a section ($1000 \leq \#D < 20000$) where the baseline curve can be approximated with steadily high accuracy using an exponential function, and the level of representativeness of W satisfying the condition of $1000 \leq \#D(W) < 20000$ is defined by a value: $Rep(W) = Dist\{PD(W), P0\} / B(\#D(W))$, that is, a value obtained by normalizing $Dist\{PD(W), P0\}$ with B(・). (It has to be noted that the "words" in this context are already cleared of all those which are considered certain to be unnecessary as query terms for information retrieval, such as symbols, particles and auxiliary verbs. While the same method can be realized even if these elements are included, in that case there will be some changes in the above-cited numerals.)

With a view to making it possible to use the well-approximated region of the aforementioned baseline function even where #D(W) is significantly great as in the case of "*suru*" and to reducing the amount of calculation, about 150 documents are extracted at random from D(W), which is denoted D'(W), so that $20,000 < \#D'(W)$ holds, and Rep(W) is calculated using D'(W) instead of D(W).

On the other hand, as the approximating function of the baseline curve figured out in the aforesaid section tends to overestimate the value in $\{x \mid 0 \leq x < 1000\}$, Rep(W) is likely to be underestimated for W in the range of $\#D(W) \leq 1000$ as a result of normalization. However, whereas 1000 words approximately correspond to two or three newspaper articles, terms which occur in the number of documents in that order is not very important for our purpose, the calculated result was applied as it was. Of course, another baseline may as well be calculated in advance. Dist{PD}, P0}/B(#D) in the randomly sampled document set D steadily gave an average, Avr, of approximately 1 (± 0.01) and a standard deviation σ of around 0.05 in various corpora. Since the maximum never surpassed $Avr + 4\sigma$, as the basis of judgment that the Rep(W) value of a given term is "a meaningful value" or not, a threshold value of $Avr + 4\sigma = 1.20$ is provided.

The above-cited measure Rep(•) has such desirable features that (1) its definition is mathematically clear, (2) it allows comparison of highly frequent terms and infrequent terms, (3) the threshold value can be defined systematically, and (4) it is applicable to terms consisting of any number of words.

The effectiveness of the measure Rep(•) proposed in the present invention has been confirmed by experiments as

well. Out of words which occurred three times or more in total in the articles in the 1966 issues of the *Nihon Keizai Shimbun*, 20,000 words were extracted at random, and 2,000 out of them were manually classified into three categories: their occurrence in the overview of retrieved contents is "desirable --- a", "neither desirable nor undesirable" and "undesirable --- d". The 20,000 words are ranked by a measure and the number of words which are classified into a specified class and appear between the first word and the Nth word, which number is hereafter called "accumulated number of words", is compared to that obtained by using another measure. In the following, four measures will be used, comprising random (i.e., no measure), frequency, tf-idf and a proposed measure. Here is used as tf-idf the version of tf-idf covering all documents, which was explained in the section on the prior art. Thus it is defined as f(w)**0.5 × log2(N/N(w)) where N is the number of all the documents, N(w) is the number of documents in which w appears, and f(w) is the frequency of w in all the documents.

Fig. 7 compares the accumulated number of words classified as "a". As is evident from the graph, the force to raise the ranks of words classified as "a" is stronger in the order of random < frequency < tf-idf < proposed measure. The improvement is evidently significant. Fig. 8 compares the accumulated numbers of words classified as

"d"; the superiority of the proposed measure in sorting capability is distinct. Frequency and tf-idf are no different from random cases, revealing their inferiority in the "stop-word" identifying capability. In view of these findings, the measure proposed according to the invention is particularly effective in identifying stop-words, and is expected to be successfully applied to the automatic preparation of a stop-word lists and the improvement of the accuracy of weighting in the calculation of document similarity by "excluding frequent but non-representative words".

An example of system configuration for the calculation of representativeness so far described is illustrated in Fig. 3. Calculation of representativeness will now be described below with reference to Figs. 3 and 4, in which 301 denotes a storage for storing document data, various programs and so forth using a hard disk or the like. It is also utilized as a working area for programs. Thereafter, 3011 denotes document data (although Japanese is used in the following example, this method is not language-specific); 3012, a morphological analysis program for identifying words constituting a document (it performs such processing as word separation by spaces and part-of-speech tagging in Japanese, or stemming in English; this method is not specified; various systems are disclosed in

both languages, whether for commercial use or research purposes); 3013, a word-document association program (for checking, according to the results of morphological analysis, which word occurs in which document and how often, or conversely in which document how many times which word occurs; basically this is a task to fill elements of a matrix having words as rows and documents as columns by counting, and no particular method is specified for this task); 3014, a word-document association database (DB) for recording word-document association data calculated as described above; 3015, a representativeness calculation program, a program for calculating the representativeness of a term, whose details are shown in Fig. 4; 3016, a DB for recording the calculated representativeness of terms; 3017, an area for a plurality of programs to reference data in a shared manner; 3018, a working area; 302, an input device; 303, a communication device; 304, a main memory; 305, a CPU; and 306, a terminal device consisting of a display, a keyboard and so forth.

Fig. 4 illustrates details of the representativeness calculation program 3015. The method of calculating the representativeness of a specific term by using this program will be described below. In the figure, 4011 denotes a module for calculating background word distribution. This module is used only once, and records the frequency of each

word in the whole documents. Thus, all words being represented by {w1, ..., wn} and Ki denoting the frequency of the occurrence of a word wi in the whole document D0 as is the case with Numerical expression 1, (K1, ... Kn) is recorded. Reference numeral 4012 denotes a module for estimating the baseline function with regard to given document data. This module, too, is used only once at the beginning. It can be realized by combining the following basic elements: (1) When the whole document sets are given, document sets the number of words in which range from around 1000 to around 20,000 are selected at random repeatedly, and at each repetition, the distance between the word distribution in each selected domument set and the word distribution in the whole documents obtained by 4011, is calculated using Numerical expression 1. (2) Baseline function B(·) is figured out using {(#D, Dist{PD, P0})}s and the least square method or the like, where D varies over randomly selected document sets in (1) and (#D, Dist{PD, P0}) was calculated for each D in (1). B(·) is a function from the number of words to a positive real number. No particular method is specified for this approximation. Standard methods are available.

Reference numeral 4013 denotes a document extraction module. When term W = wn1 ... wnk is given, a document set D(wni)($1 \leq i \leq k$) is obtained from the word-document

association DB 3014 and the intersection of all D(wni) ($1 \leq i \leq k$) is taken to determine D(W). If the word-document association DB 3014 records the information on the position of a word in every document, the set of all documents containing term W = wn1 ... wnk can be obtained, which is a subset of the intersection of all D(wni) ($1 \leq i \leq k$). If the word-document association DB 3014 does not record the information on the position of a word in the document, the intersection of all D(wni) ($1 \leq i \leq k$) is taken as D(W) as an approximation. Numeral 4014 denotes a module for calculating co-occurring word distribution. Again the frequency of each word in D(W) is counted from the word-document association DB 3014 to determine the frequency ki of wi in D(W) ($1 \leq i \leq k$). Numeral 4015 denotes a module for calculating distance between word distributions. Using Numerical expression 1 and the word frequencies obtained by 4011 and 4014, the distance Dist{PD(W), P0} between the word distribution in the whole documents and the word distribution in D(W) is calculated. Numeral 4016 denotes a module for normalizing the aforementioned distance Dist{PD(W), P0}. Using the number of words in D(W), which is denoted #D(W), and B($\cdot$) obtained by 4012, it calculates the representativeness of W as Rep(W)= Dist{PD(W), P0}/B(#D(W)). Numeral 4017 denotes a random sampling module, which is used in 4013 to select a predetermined

number of documents when the number of documents contained in D(W) surpasses a predetermined number (recorded in the shared data area 3017). While in this instance the number of documents is used as the predetermined number, it is also possible to use the desirable number of words as the predetermined number and to make the number of words in randomly sample documents as close to the predetermined number as possible.

Fig. 5 shows an example of configuration for the application of the invention for assisting document retrieval. This diagram illustrates the configuration of a retrieving apparatus where the invention is applied to the displaying of topic words in a navigation window in line with the configuration shown in Fig. 1 of the document retrieval support method according to Reference 1. It differs from the document retrieval support method according to Reference 1 in that, in a topic words displaying routine 544, a representativeness check routine 5445 is added, and in a topic words extraction routine 5441, a co-occurrence analysis routine 5442, a graph mapping routine 5443 and a graph displaying routine 5444, the representativeness check routine is used. The representativeness check routine is a routine to return the representativeness of each word in the set of the whole documents. It is possible to calculate

in advance the representativeness of each word according to the program shown in Fig. 4.

When the user enters a retrieval keyword from a keyboard 511, the titles of the documents containing that keyword, which are the result of retrieval, are displayed on a user-interface window for information retrieval 521, and topic words selected out of the document set are displayed on a window for displaying topic words 522. First, words are selected in the topic words extraction routine 5441 by the method of Reference 1. Although the word selected here include, as stated earlier, common words such as "*suru*" and "*kono*" (this), the displaying of highly frequent stop-words can be suppressed by checking the representativeness of words according to the representativeness check routine 5445 and excluding words whose representativeness values are smaller than a preset threshold (for instance, 1.2). Furthermore, if displayed words overlap each other by the method of Reference 1, it is easy to display more to the front the word higher in representativeness or to display in heavier tone the word higher in representativeness by using the representativeness check routine 5445 in the graph mapping routine 5443 and the graph displaying routine 5444. Thus it is possible to display words higher in representativeness in a more conspicuous way and thereby improve the user

interface. Furthermore, while the foregoing description suggested calculation of the representativeness of each word in advance according to the program shown in Fig. 4, it is also possible to regard each set of documents obtained for each input keyword as the set of whole documents anew, and calculate according to the program shown in Fig. 4 the representativeness of each word contained in the documents, which is the result of retrieval, as it occurs. If the representativeness check routine 5445 is so designed, the representativeness of the same word may differ with the keyword, and accordingly it will be possible to display topic words in a manner reflecting the retrieval situation more appropriately.

Fig. 6 shows an example of configuration for use in applying representativeness to automatic word extraction. In the figure, 601 denotes a storage for storing document data, various programs and so forth using a hard disk or the like. It is also utilized as a working area for programs. Thereafter, 6011 denotes document data (although Japanese is used in the following example, this method is not language-specific); 6012, a morphological analysis program for identifying words constituting a document (it performs such processing as word separation by spaces and part-of-speech tagging in Japanese, or stemming in English; this method is not specified; various systems are disclosed in

both languages, whether for commercial use or research purposes); 6013, a word-document association program (for checking, according to the results of morphological analysis, which word occurs in which document and how often, or conversely in which document how many times which word occurs; basically this is a task to fill elements of a matrix having words as rows and documents as columns by counting, and no particular method is specified for this task); 6014, a word-document association database (DB) for recording word-document association data calculated as described above; 6015, an extracted word storing DB; 6017, a representativeness calculation program, whose details are shown in Fig. 4; 6018, a program for calculating the representativeness of a term; 6019, an area for a plurality of programs to reference data in a shared manner; 601A, a program to select the words or word sequences which will become the candidates for extraction (though the contents are not specified, words such as particles, auxiliary verbs and affixes are usually excluded from a given result of document morphological analysis); 601B, a filter for utilizing grammatical knowledge to exclude word sequences unsuitable as terms out of the candidates selected by 601A (for instance, sequences in which a case affix or an auxiliary verb comes first or last are excluded; though the contents are not specified, a number of examples are

mentioned in the paper cited as Reference 2). The candidates selected by 601B undergo the calculation of importance by 601C according to a specific measure and, those lower than a preset level of that measure being excluded, are sorted according to importance and outputted. While this is called the tf_idf filter program after the name of the most frequently used measure, the actually used measure may be any appropriate measure other than tfi_df. Reference numeral 6016 denotes a working area; 602, an input device; 603, a communication device; 604, a main memory; 605, a CPU; and 606, a terminal device consisting of a display, a keyboard and so forth. The usual word extraction method uses neither 6017 nor 6018. In response to the output of 601C, the representativeness of each candidate is referenced by 6017 and 6018, and those whose measures are lower than a preset level (for instance 1.2) are excluded. A conceivable variation would use 6017 and 6018 in 601C to directly reference the representativeness of each candidate, and select the candidate terms according to representativeness as the sole criterion.

An experiment was carried out using the automatic word extraction method of the configuration illustrated in Fig. 6, and terms were extracted from the abstracts of 1,870 papers on artificial intelligence. About 18,000 term candidates were extracted by 601A and 601B. Two procedures

were tested: in one procedure only representativeness was used and in the other term candidates were first sorted by tf-idf and the output of sorting was cleared of unimportant words by using representativeness. The two procedures equally gave about 5,000 term candidates, but the latter tended to extract terms in a sequence close to the order of frequency, so that in seeking final selection by human judgment, the latter may be more natural in a way because familiar words come relatively early.

By using representativeness as proposed in the present invention, there is provided a representativeness calculation which, with respect to terms in a document set, (1) gives a clear mathematical meaning, (2) permits comparison of high-frequency terms and low-frequency terms, (3) makes possible setting of a threshold value in a systematic way, and (4) is applicable to terms containing any number of words. Thus a method to calculate the importance of words or word sequences can be realized, which would prove useful in improving the accuracy of word information retrieval interfaces and word extraction systems.

While the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that the foregoing and other changes in form and details can be made

therein without departing from the spirit and scope of the invention.

WHAT IS CLAIMED IS:

1. A word importance calculation method for calculating the importance of words contained in a document set, whereby the difference between the word distribution in a subset of whole documents which consists of every document containing a specified word and the word distribution in the set of whole documents is used to calculate the importance of the word.

2. A word importance calculation method, as claimed in Claim 1, wherein:

said difference is determined by comparing the distance d between said subset and said set of whole documents with the distance d', or the estimated value of d', between another subset of documents which contain substantially the same number of words as said subset of documents and are randomly selected from said set of whole documents, and said set of whole documents.

3. A word importance calculation method, as claimed in Claim 2, wherein:

the distance d between the two document sets is calculated by using the word distribution in each document set, that is to say using the probability of occurrence of each word in each of said document set.

4. A word importance calculation method, as claimed in Claim 2, wherein:

if the number of documents containing said word is larger than a prescribed number, a preset number of documents are extracted from the said subset of whole documents by random sampling, and the difference between the extracted set of documents and said set of whole documents is used instead of the difference between the original subset of documents and the set of whole documents.

5. A document retrieval interface having a function to display on a screen words characterizing a document set, wherein the importance of each word occurring in the set of whole documents is calculated using the difference between the word distribution in the subset of whole documents containing the word and the word distribution in the set of whole documents, and the importance is brought to bear on the selection, arrangement or coloring of the words displayed on the screen.

6. A document retrieval interface having a function to display on a screen words characterizing a document set, wherein the importance of each word occurring in the document set obtained as a result of retrieval is calculated using the difference between the word distribution in the subset of documents out of the document set obtained as a result of that retrieval containing that word and the word distribution in the document set obtained as a result of that retrieval, and the importance is brought to bear on the

selection, arrangement or coloring of the words displayed on the screen.

7. A word dictionary construction method by extracting important words from a document set in accordance with rules given in advance, wherein the importance of each word occurring in a set of whole documents is calculated using the difference between a subset of whole documents containing the word and the word distribution in the set of whole documents, and words to be extracted are selected on the basis of that importance.

8. A word importance calculation method whereby;

a characteristic quantity of a document set containing a certain word and

a characteristic quantity of a randomly extracted document set of the size of said set are compared; and

the importance of said word is thereby calculated.

9. A word importance calculation method, as claimed in Claim 8, wherein:

the difference between the word distribution in said document set and the word distribution in the set of all documents is used as said characteristic quantity.

ABSTRACT OF THE DISCLOSURE

A method according to the prior art for selecting words (or word sequences), which is an important aspect of information retrieval, involves the problems of inability to eliminate high-frequency common words and of often arbitrary setting of the threshold value for dividing important and unimportant words. These problems are solved by normalizing the difference between the word distribution in a subset of all documents containing a word to be extracted (or a subset of said document set) and the word distribution in the set of whole documents with the number of words in the said subset of whole documents containing the word as a parameter, and the accuracy of support to information retrieval is thereby enhanced.

FIG. 1

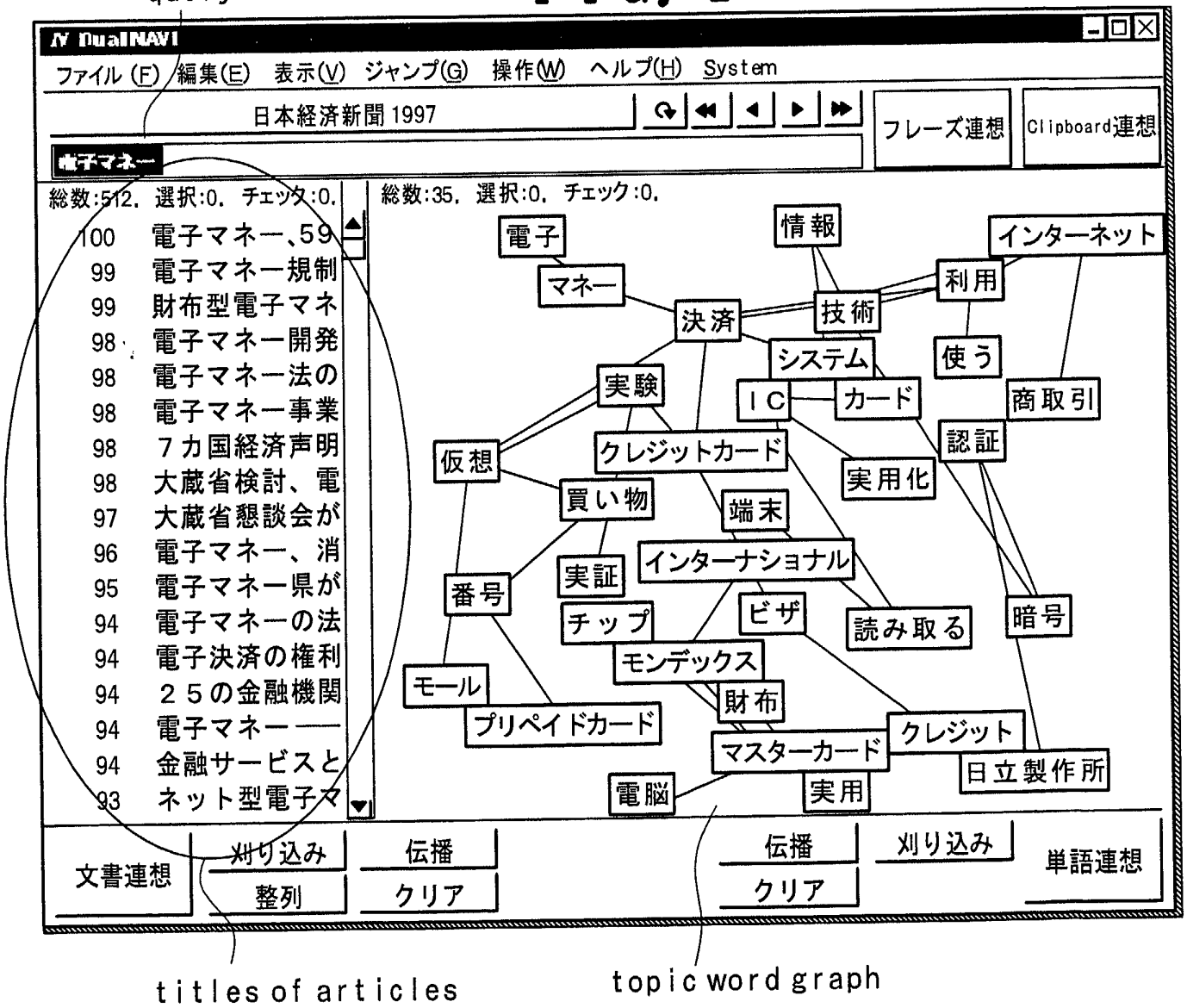


FIG. 2

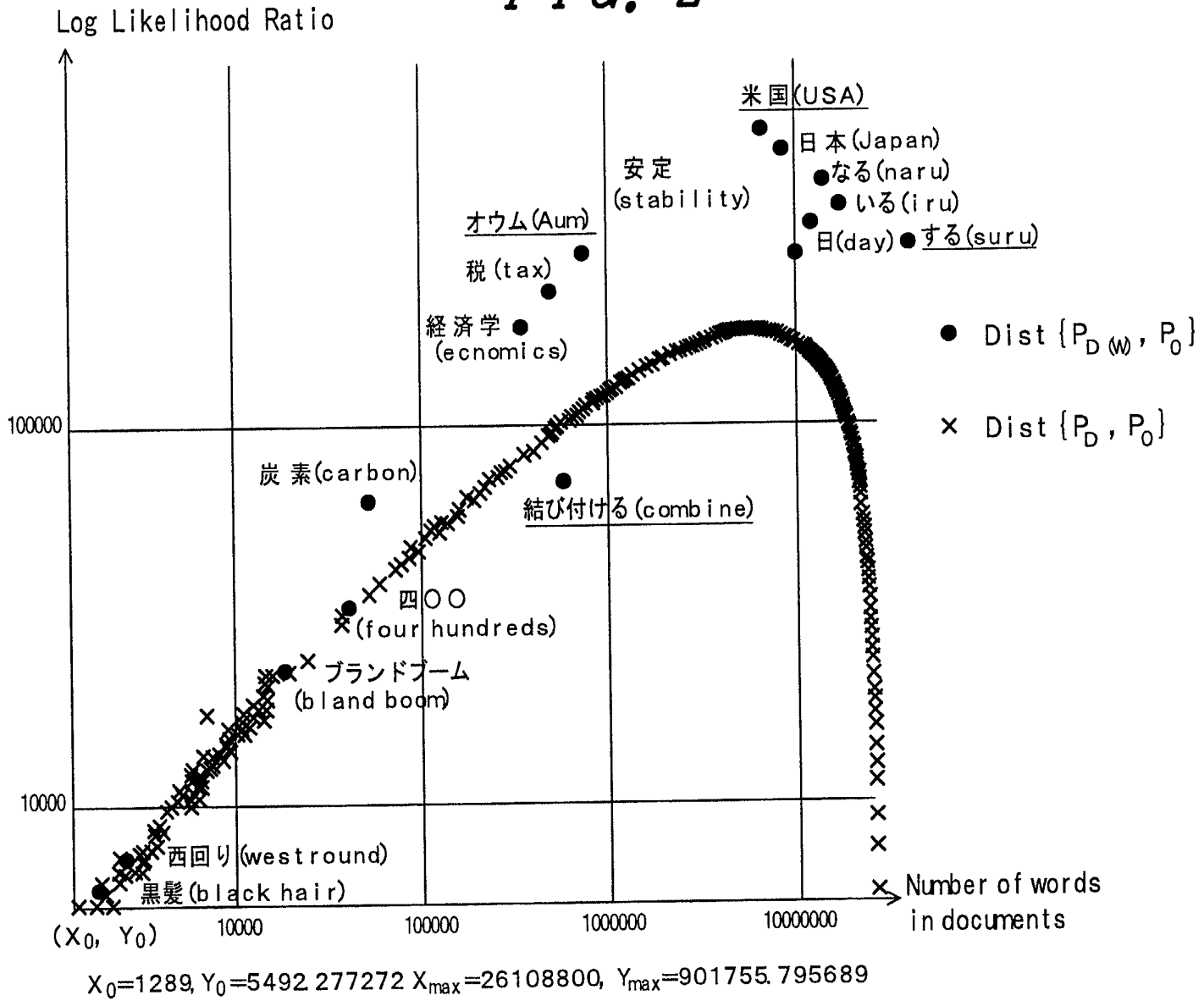

Log Likelihood Ratio
米国(USA)
日本(Japan)
なる(naru)
いる(iru)
日(day)
する(suru)
安定
(stability)
オウム(Aum)
税(tax)
経済学
(ecnomics)
Dist {P_D(W), P_0}
Dist {P_D, P_0}
100000
炭素(carbon)
結び付ける(combine)
四〇〇
(four hundreds)
ブランドブーム
(bland boom)
10000
西回り(westround)
黒髪(black hair)
(X_0, Y_0)
10000
100000
1000000
10000000
Number of words
in documents
X_0=1289, Y_0=5492.277272 X_max=26108800, Y_max=901755.795689

FIG. 3

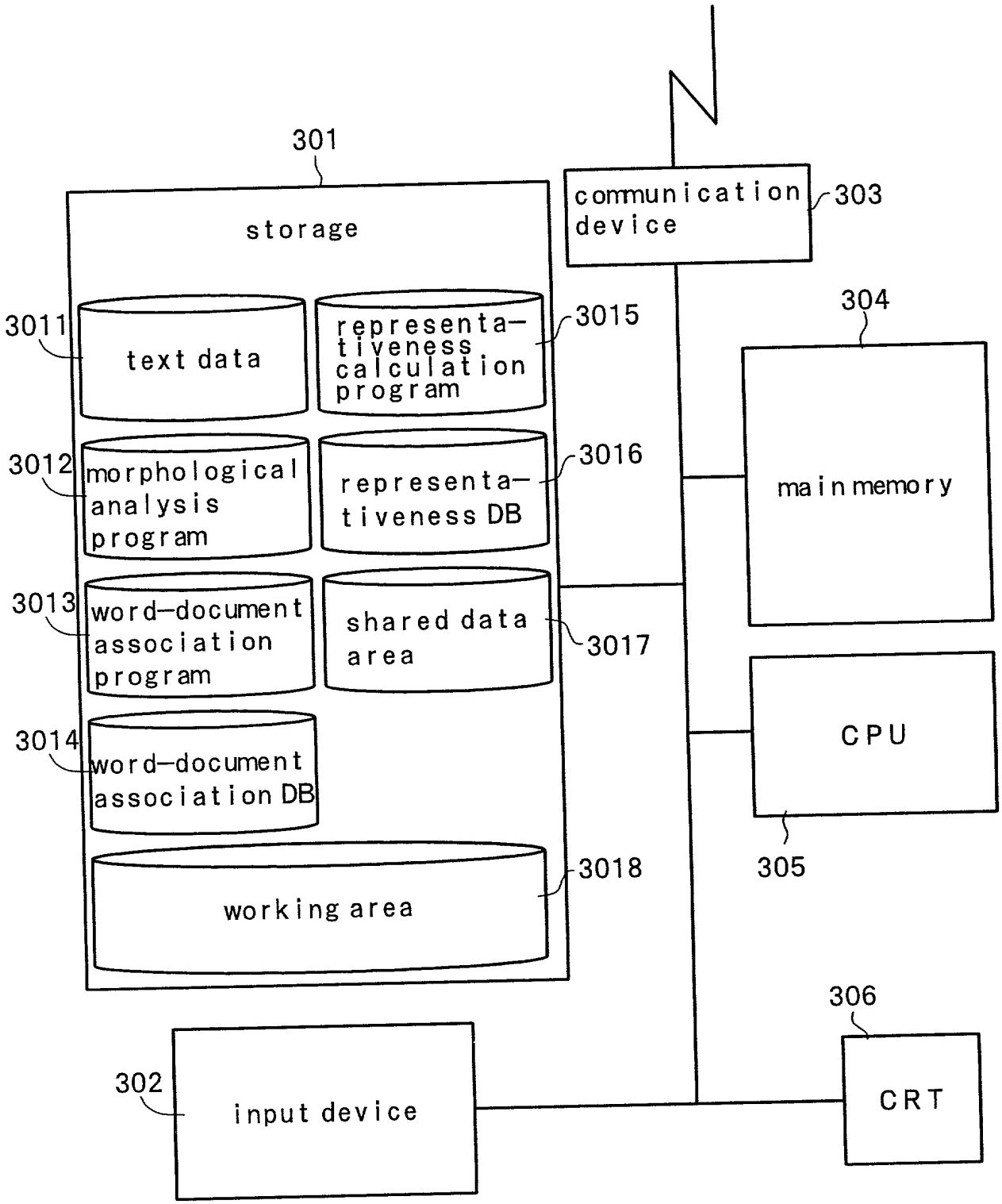

301
storage
communication device
303
304
3011
text data
representa-tiveness calculation program
3015
3012
morphological analysis program
representa-tiveness DB
3016
main memory
3013
word-document association program
shared data area
3017
3014
word-document association DB
CPU
305
working area
3018
306
302
input device
CRT

FIG. 4

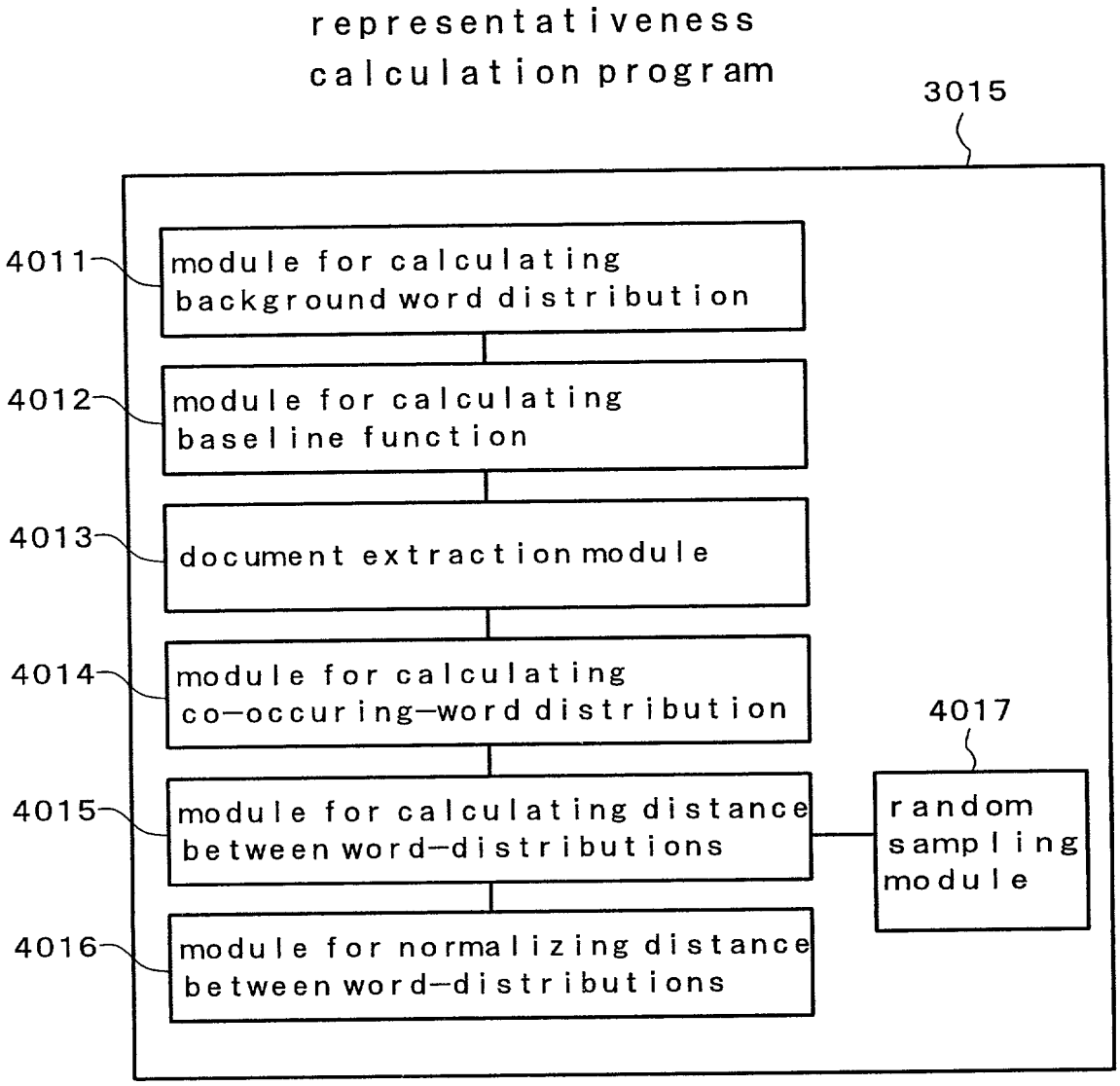

representativeness calculation program
3015
4011
module for calculating background word distribution
4012
module for calculating baseline function
4013
document extraction module
4014
module for calculating co-occuring-word distribution
4015
module for calculating distance between word-distributions
4016
module for normalizing distance between word-distributions
4017
random sampling module

FIG. 5

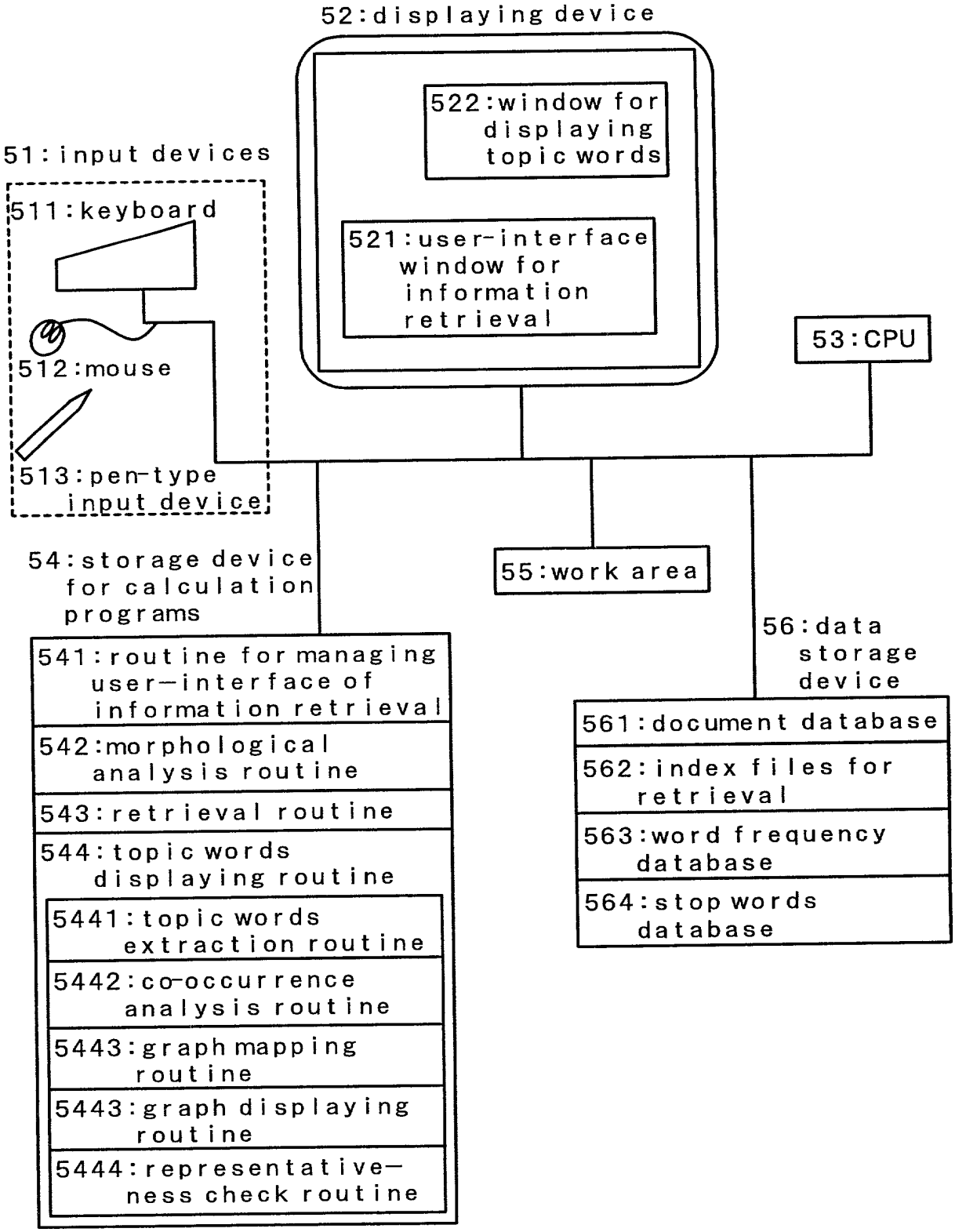

52:displaying device
522:window for displaying topic words
521:user-interface window for information retrieval
51:input devices
511:keyboard
512:mouse
513:pen-type input device
53:CPU
54:storage device for calculation programs
55:work area
56:data storage device
541:routine for managing user-interface of information retrieval
542:morphological analysis routine
543:retrieval routine
544:topic words displaying routine
5441:topic words extraction routine
5442:co-occurrence analysis routine
5443:graph mapping routine
5443:graph displaying routine
5444:representative-ness check routine
561:document database
562:index files for retrieval
563:word frequency database
564:stop words database

FIG. 6

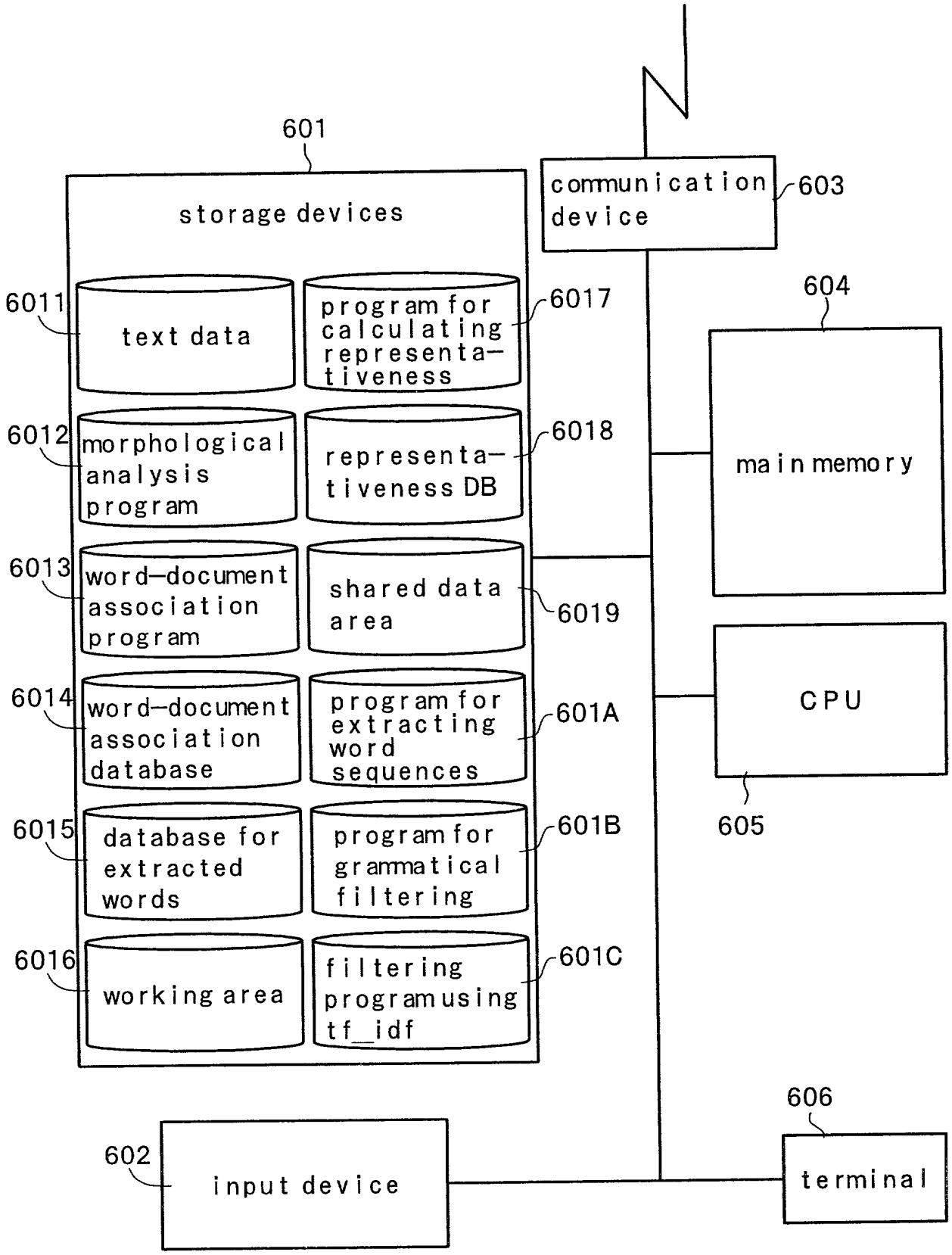

601
storage devices
6011
text data
6012
morphological analysis program
6013
word–document association program
6014
word–document association database
6015
database for extracted words
6016
working area
program for calculating representa–tiveness
6017
representa–tiveness DB
6018
shared data area
6019
program for extracting word sequences
601A
program for grammatical filtering
601B
filtering programusing tf_idf
601C
communication device
603
604
main memory
CPU
605
606
terminal
602
input device

FIG. 7

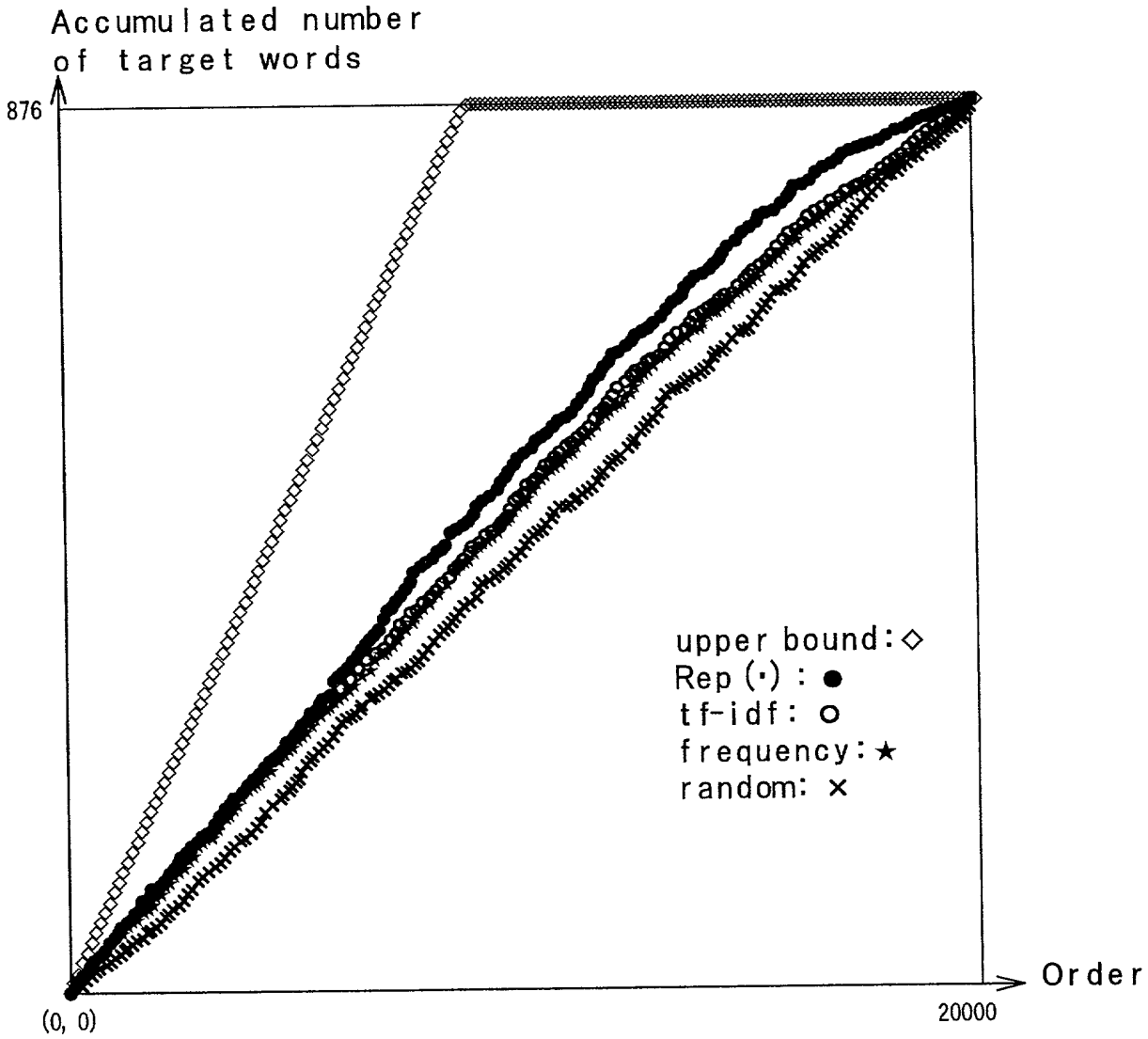
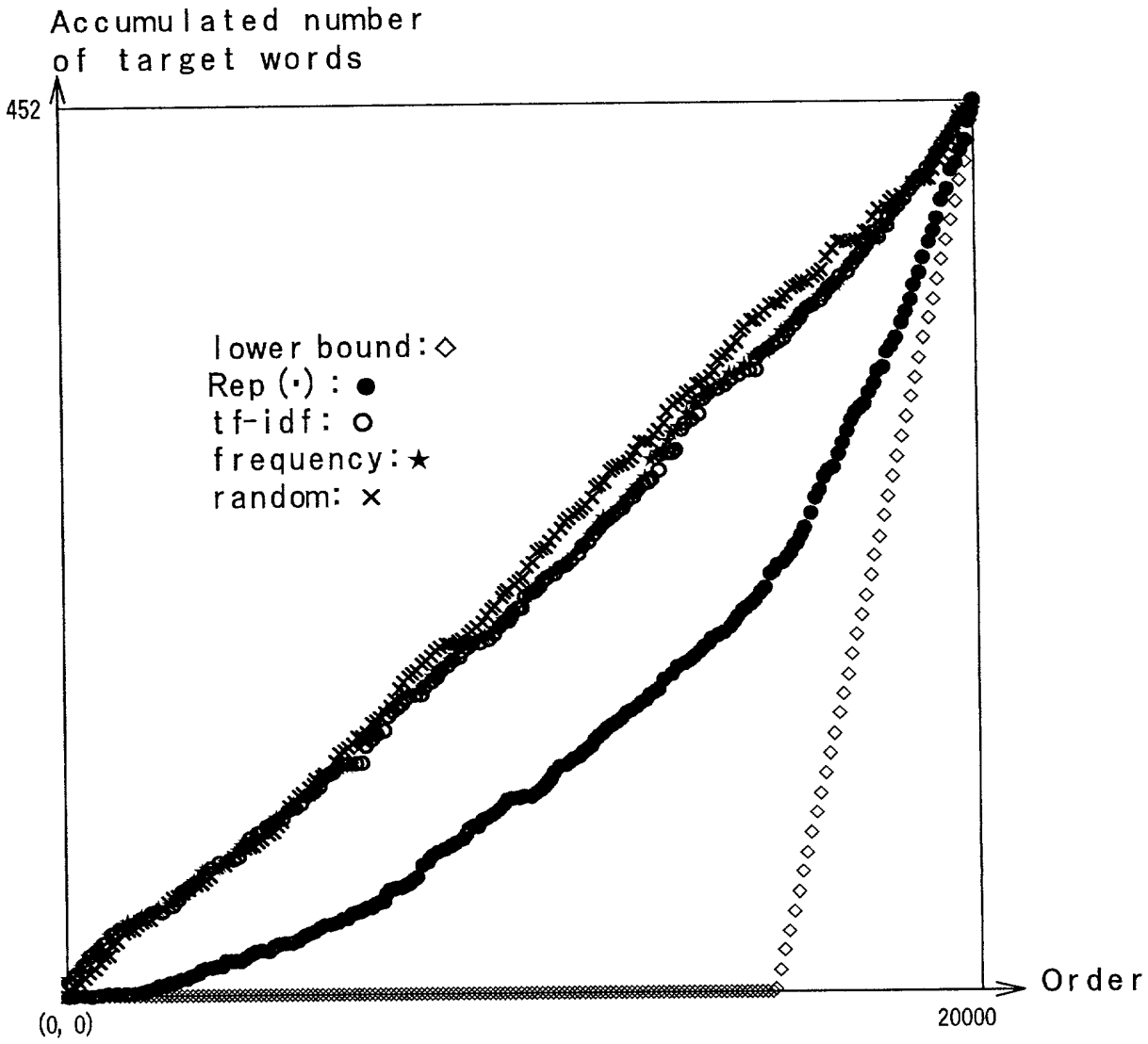

Accumulated number
of target words
876
(0, 0)
20000
Order
upper bound: ◇
Rep (·) : ●
tf-idf: o
frequency: ★
random: ×

FIG. 8


Accumulated number
of target words
452
lower bound: ◇
Rep (·) : ●
tf-idf: o
frequency: ★
random: ×
Order
(0, 0)
20000

PTO/SB/106(8-96)
Approved for use through 9/30/98 OMB 0651-0032
Patent and Trademark Office, U S DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

Declaration and Power of Attorney For Patent Application

特許出願宣言書及び委任状

Japanese Language Declaration

日本語宣言書

下記の氏名の発明者として、私は以下の通り宣言します。

As a below named inventor, I hereby declare that:

私の住所、私書箱、国籍は下記の私の氏名の後に記載された通りです。

My residence, post office address and citizenship are as stated next to my name.

下記の名称の発明に関して請求範囲に記載され、特許出願している発明内容について、私が最初かつ唯一の発明者（下記の氏名が一つの場合）もしくは最初かつ共同発明者であると（下記の名称が複数の場合）信じています。

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

WORD IMPORTANCE CALCULATION METHOD,
DOCUMENT RETRIEVING INTERFACE, WORD
DICTIONARY MAKING METHOD

上記発明の明細書（下記の欄で×印がついていない場合は、本書に添付）は、

The specification of which is attached hereto unless the following box is checked

☐ __月__日に提出され、米国出願番号または特許協定条約国際出願番号を_____とし、
(該当する場合) _____に訂正されました。

☐ was filed on _____
as United States Application Number or
PCT International Application Number
_____ and was amended on
_____ (if applicable).

私は、特許請求範囲を含む上記訂正後の明細書を検討し、内容を理解していることをここに表明します。

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

私は、連邦規則法典第３７編第１条５６項に定義されるとおり、特許資格の有無について重要な情報を開示する義務があることを認めます。

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56

PTO/SB/106(8-96)
Approved for use through 9/30/98 OMB 0651-0032
Patent and Trademark Office, U S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

Japanese Language Declaration
(日本語宣言書)

私は、米国法典第３５編１１９条 (a)-(d) 項又は３６５条(b)項に基き下記の、米国以外の国の少なくとも一カ国を指定している特許協力条約３６５ (a) 項に基ずく国際出願、又は外国での特許出願もしくは発明者証の出願についての外国優先権をここに主張するとともに、優先権を主張している、本出願の前に出願された特許または発明者証の外国出願を以下に、枠内をマークすることで、示しています。

I hereby claim foreign priority under Title 35, United States Code, Section 119 (a)-(d) or 365(b) of any foreign application(s) for patent or inventor's certificate, or 365(a) of any PCT international application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or PCT International application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)
外国での先行出願

(Number) (番号)	(Country) (国名)	(Day/Month/Year Filed) (出願年月日)	Priority Not Claimed 優先権主張なし
11-237845	Japan	25 / August / 1999	☐
			☐

私は、第３５編米国法典１１９条 (e) 項に基いて下記の米国特許出願規定に記載された権利をここに主張いたします。

I hereby claim the benefit under Title 35, United States Code, Section 119(e) of any United States provisional application(s) listed below.

(Application No.) (出願番号)	(Filing Date) (出願日)	(Application No.) (出願番号)	(Filing Date) (出願日)

私は、下記の米国法典第３５編１２０条に基いて下記の米国特許出願に記載された権利、又は米国を指定している特許協力条約３６５条 (c) に基ずく権利をここに主張します。また、本出願の各請求範囲の内容が米国法典第３５編１１２条第１項又は特許協力条約で規定された方法で先行する米国特許出願に開示されていない限り、その先行米国出願書提出日以降で本出願書の日本国内または特許協力条約国際提出日までの期間中に入手された、連邦規則法典第３７編１条５６項で定義された特許資格の有無に関する重要な情報について開示義務があることを認識しています。

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s), or 365(c) of any PCT international application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application in the manner provided by the first paragraph of Title 35, United States Code Section 112, I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of application.

(Application No.) (出願番号)	(Filing Date) (出願日)	(Status: Patented, Pending, Abandoned) (現況：特許許可済、係属中、放棄済)

私は、私自身の知識に基ずいて本宣言書中で私が行なう表明が真実であり、かつ私の入手した情報と私の信じるところに基ずく表明が全て真実であると信じていること、さらに故意になされた虚偽の表明及びそれと同等の行為は米国法典第１８編第１００１条に基ずき、罰金または拘禁、もしくはその両方により処罰されること、そしてそのような故意による虚偽の声明を行なえば、出願した、又は既に許可された特許の有効性が失われることを認識し、よってここに上記のごとく宣誓を致します。

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

PTO/SB/106(8-96) (Modulated spacing)
Approved for use through 9/30/98. OMB 0651-0032
Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

Japanese Language Declaration
(日本語宣言書)

委任状： 私は下記の発明者として、本出願に関する一切の手続きを米特許商標局に対して遂行する弁理士または代理人として、下記の者を指名いたします。（弁護士、または代理人の氏名及び登録番号を明記のこと）

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith (*list name and registration number*)

Donald R. Antonelli, Reg. No. 20,296; David T. Terry, Reg. No. 20,178; Melvin Kraus, Reg No. 22,466; William I. Solomon, Reg. No. 28,565; Gregory E. Montone, Reg. No. 28,141; Ronald J. Shore, Reg. No. 28,577; Donald E. Stout, Reg. No. 26,422; Alan E. Schiavelli, Reg. No. 32,087; James N. Dresser, Reg. No. 22,973 and Carl I. Brundidge, Reg. No. 29,621

書類送付先

Send Correspondence to:
Antonelli, Terry, Stout & Kraus, LLP
Suite 1800
1300 North Seventeenth Street
Arlington, Virginia 22209

直接電話連絡先： (名前及び電話番号)

Direct Telephone Calls to: (*name and telephone number*)

Telephone: (703) 312-6600
Fax: (703) 312-6666

唯一または第一発明者名	**Full name of sole or first inventor** Toru HISAMITSU
発明者の署名 日付	**Inventor's signature** 久光 徹 **Date** 6/21/2000
住所	**Residence** Oi, Japan
国籍	**Citizenship** Japan
私書箱	**Post Office Address** c/o Hitachi, Ltd., Intellectual Property Group New Marunouchi Bldg. 5-1, Marunouchi 1-chome, Chiyoda-ku, Tokyo 100-8220, Japan

(第二以降の共同発明者についても同様に記載し、署名をすること)

(Supply similar information and signature for second and subsequent joint inventors.)

第二共同発明者名		**Full name of second joint inventor, if any** Yoshiki NIWA
第二共同発明者の署名	日付	**Second inventor's signature** **Date** 丹羽芳樹 6/21/2000
住所		**Residence** Hatoyama, Japan
国籍		**Citizenship** Japan
私書箱		**Post Office Address** c/o Hitachi, Ltd., Intellectual Property Group New Marunouchi Bldg. 5-1, Marunouchi 1-chome, Chiyoda-ku, Tokyo 100-8220, Japan
第三共同発明者名		**Full name of third joint inventor, if any**
第三共同発明者の署名	日付	**Third inventor's signature** **Date**
住所		**Residence**
国籍		**Citizenship**
私書箱		**Post Office Address**
第四共同発明者名		**Full name of fourth joint inventor, if any**
第四共同発明者の署名	日付	**Fourth inventor's signature** **Date**
住所		**Residence**
国籍		**Citizenship**
私書箱		**Post Office Address**
第五共同発明者名		**Full name of fifth joint inventor, if any**
第五共同発明者の署名	日付	**Fifth inventor's signature** **Date**
住所		**Residence**
国籍		**Citizenship**
私書箱		**Post Office Address**